ROMANIAN JOURNAL OF METEOROLOGY
Volume 11
Issues 1-2
2014

Calibration of extreme temperature forecasts of MOS_EPS model over Romania with the Bayesian Model Averaging

Mihaela-Silvana NEACSU
National Meteorological Administration , Bucharest
E-mail: mihaela.ridichie@meteoromania.ro

Abstract. The paper presents the use of the BMA - Bayesian Model Averaging method to calibrate the forecasts of extreme temperatures obtained from the model MOS_EPS for the local scale forecast offered by ECMWF, with anticipations of up to 15 days. The theoretical aspects of BMA and the manner in which this technique has been applied for MOS_EPS, the results obtained for a few weather stations and the verification methods for the performances of the BMA model are also described in this paper.
Results show that the probability density function obtained through the implementation of the BMA method was much better calibrated than the probability density functions of the each members of the ensemble. The mean square error of the ensemble calibrated with BMA was lower than the mean square error of the MOS_EPS ensemble in 64% of the studied cases and lower in comparison to the mean square error of each member of the calibrated ensemble in 35% of the cases.

Keywords: ensemble forecasting, calibration, BMA, extreme temperature, Romania.

1. INTRODUCTION

Present-day EPS (The Ensemble Prediction System) techniques consist in integrating a deterministic forecast with an estimation of the function of probability distribution of forecast states. MOS_EPS (Model Output Statistics Ensemble Prediction System) - is a model of statistical adjustment implemented since 2008 in the National Meteorological Administration of Romania. This model is used for extreme temperatures deterministic point forecasting, using the forecasting ensemble ECMWF (European Center for Medium range Weather Forecasting) for anticipations of up to 15 days.

The purpose of the method is using the information regarding the discrepancy of numerical solutions obtained for small perturbations in the initial conditions, to the limit or in physical parameterizations, to create the ensemble forecast as well as to detect those subdomains of parameter space with increases sensitivity to perturbations. Optimizing the measurements system - especially in these subdomains, increases the performance of the numerical solution.

In recent years, several statistical methods to calibrate the forecasts offered by the members of the forecasting ensemble have been used, such as: logistic regression (Wilks, 2006), the Bayesian Model Averaging method – BMA (Raftery et al., 2005 and Fraley et al., 2011a), non-homogeneous Gaussian regression (Gneiting et al., 2005) and the Gaussian ensemble dressing (Roulston and Smith, 2003; Wang and Bishop, 2005). In this paper the BMA techniques proposed by Raftery et al. (2005) are employed to calibrate the extreme air temperature in MOS_EPS at 163 stations in Romania. The results presented are those obtained for Băneasa, Constanţa, Craiova, Sibiu, Bacău and Timişoara stations.

For temperature, a mixture of Gaussian distributions is used. The output parameters after applying the BMA method are estimated from a test archive which contains forecasts and observations. The estimation of parameters is completed by maximizing the log likelihood (maximum likelihood estimator application) or by minimizing the CRPS (Continuous Ranked Probability Score).

BMA was initially developed for weather parameters whose PDF (Probability Density Function) can be approximated by a normal distribution (temperature and sea-level pressure), then the application became applicable to parameters whose PDFs don't have a distribution which can be approximated with a Gaussian one, such as precipitation and wind direction.

2. DATA AND METHODS

For this study, we used an initial archive of 60 days, starting with the 1st of August 2012. The archive contains forecast data of MOP_EPS extreme temperatures, observations, identification code and validity data of the weather station. This archive has been used as test period.

MAE – Mean Absolute Error is a measurement of the accuracy of forecasts and is calculated according to the formula:

$$MAE = \frac{1}{N}\left(\sum_{i=1}^{N} |F_i - O_i|\right)$$

MAE shows the average amplitude error, but does not indicate the meaning of the bias.

CRPS is calculated using the formula:

$$CRPS = \frac{1}{n}\sum_{st}\int_{-\infty}^{+\infty}\left(H_{st}(y) - 1_{\{y \geq \Delta_{st}\}}\right)^2 dy$$

where H(y) represents the cumulative distribution of BMA and predictive probability density function, $1_{\{y \geq \Delta_{st}\}}$ represents the Hessian function

$$1_{\{y \geq \Delta_{st}\}} = \begin{cases} 1, if\ 1_{\{y \geq \Delta_{st}\}} \\ 0, \quad otherwise \end{cases}$$

and *n* is the total number of observations.

CRPS is equivalent with the mean square error for a deterministic forecast (Hersbach, 2000). The lower values of these measures indicate a better performance of the predictive BMA model.

The results of the MAE and CRPS functions for each anticipation day for the 163 stations are presented in Table 1 and Table 2 respectively.

Table 1. Mean Absolute Error (MAE) for anticipations 24, 48, 72 and respectively 96 hours, for test periods of 10, 15, 20, 25, 30, 35 and respectively 40 days.

	10days	15days	20days	25days	30days	35days	40days
24 h	4.726127	4.716683	4.783942	4.806286	4.7382269	4.6256810	4.545242
48 h	4.963846	4.917804	4.926387	4.810671	4.7672954	4.6321562	4.625424
72 h	5.038628	4.251730	4.119127	3.901796	4.2785213	4.2163402	4.719360
96 h	4.311004	4.240217	4.199544	4.172472	4.525996	4.389911	4.420161

Table 2. Continuous Ranked Probability Score (CRPS) for anticipations 24, 48, 72 and respectively 96 hours, for test periods of 10, 15, 20, 25, 30, 35 and respectively 40 days.

	10days	15days	20days	25days	30days	35days	40days
24 h	5.379370	5.441745	5.505655	5.491021	5.407119	5.228851	5.081497
48 h	5.459410	5.494700	5.560186	5.503479	5.414839	5.228700	5.139034
72 h	5.489741	4.836809	4.418327	4.331528	4.737278	4.448421	4.567887
96 h	4.594666	4.4919867	4.28044	4.418993	4.517185	4.515230	4.558307

Analyzing the results of the simulation of the BMA method with "learning" periods of 10, 15, 20, 25, 30, 35 and respectively 40 days with anticipations of 24, 48, 72 and respectively 96 hours, it has been established that the number of days which lead to a better performance, in terms of time and resources, should be 35 days, considering that this conclusion has been established using an archive which contains 50 days with EPS forecasts. Each day contains the forecasts of the 51 integrations of the MOS_EPS model (lowest temperatures separated from the highest), for each of the 163 stations (each day predicted with up to 14 days anticipation).

For the evaluation of performance, we used the mean square error calculated as follows:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(F_i - O_i)^2}$$

where N is a series of forecasts. F_i represents the predicted value, and O_i the observed value.

The RMSE formula for the ensemble of predictions calibrated with the BMA method:

$$RMSE = \sqrt{\frac{\sum_{s,t}\left(\sum_{k=1}^{K} w_k f_k - \Delta_{st}\right)^2}{N}}$$

The RMSE and ensemble mean values for the 2nd of October with 14 days of anticipation, calculated for maximum temperature before and after calibration, are shown in Table 3.

Table 3. Root Mean Square Error for the calibrated ensemble, un-calibrated and the calibrated ensemble mean respectively.

RMSE after calibration	RMSE befor calibration	Ensemble mean after calibration
9.330033	9.131479	6.663081
8.822632	8.84286	6.949275
12.66031	12.88869	7.399684
16.37765	16.22168	6.529851
12.86918	12.12856	4.719406
8.677081	8.876206	4.811453
8.877281	9.017714	5.126952
12.68872	12.75431	5.566209
13.02456	13.32581	5.660484
12.91104	13.1829	5.796534
16.03761	16.31507	5.657815
9.697504	9.986947	5.467481
9.47348	9.433724	5.554447
10.38907	10.08602	5.527543

The distribution of observation data is summarized with the help of histograms. Starting with a separation in class intervals, the histogram represents a series of rectangles having as bases the class intervals and the areas proportional with the number of observations belonging to class intervals. The height of a rectangle is calculated as the proportion between

the number of observations in the class interval and the length of the respective interval. The vertical axis of a histogram is thus a density scale. In the construction of a histogram, the endpoint convention is important.

The Verification rank histogram is constructed as follows. The 51 predictions in ascending order (R_1, …, R_{51}) define 52 intervals (50 intervals between the members of the predictions and two outside them). From this diagram, we can notice the observation rank for each day. The ideal Rank checker histogram is the one in which the frequency of observations is equal to 1 for all members, meaning we have the equiprobability hypothesis. In general, we encounter U-shaped Rank histograms, in which the dispersion function of the forecast ensemble is not very high and the majority of the observations are in extreme intervals.

As the extreme temperatures are not limited, the endpoints need to be defined.

Let f_i be the frequency for observations for the [R_{i-1}, R_i] interval. The cumulative frequency of the distribution functions noted with F_i depends on f_i thus:

$$F_i = \sum_{j=1}^{i} f_j \, , i = 1, \dots, 52$$

The BMA method – Bayesian Model Averaging described by Raftery et al. (2005) and Jasper et al. (2006):

Let Δ be the parameter to be forecast and Δ^t the set of training data for the test period using K statistical models (M_1,…, M_K). The total probability function shows that the probability density function (PDF) - $p(\Delta)$ is given by the following formula:

$$p(\Delta) = \sum_{k=1}^{K} p(\Delta|\mathrm{M_k}) p(M_k|\Delta^T)$$

where $p(\Delta|\mathrm{M_k})$ is the forecast PDF based only on the model M_K , and $p(M_k|\Delta^T)$ is the posterior probability of model M_K being correct given the training data, and reflects how well model Mk fits the training data. The posterior model probabilities add up to one, meaning:

$$\sum_{k=1}^{K} p\left(M_k\middle|\Delta^T\right) = 1$$

resulting that these can be seen as weights.

The PDF of the BMA model is a weighted mean of the conditioned PDFs given by each model, themselves weighed by the posterior probabilities of the model.

Be f as $\{f_1, \ldots, f_K\}$ which describes the ensemble forecast obtained from K different models and be Δ the quantity of interest. Through the BMA method each member of the forecast ensemble is associated with a conditional probability density function, $g_k(\Delta \mid f_k)$ which can be interpreted as the PDF of Δ given by f_K. The BMA predictive model for the dynamic ensemble forecast can be expressed as:

$$p(f_1, \ldots, f_K) = \sum_{k=1}^{K} w_k \, g_k(\Delta \mid f_k)$$

where w_k is the posterior probability of forecast K (the best one). w_k have nonnegative values and added up they reach 1, and they can thus be viewed as weights, the weights of each model in relation to the whole ensemble and they show the contribution of the relative model to the relative accuracy towards the training period.

The original method of the BMA ensemble described by Raftery et al. (2005) implies that the conditional PDFs $g_k(\Delta \mid f_k)$ of the forecast ensemble members can be approximated with a normal distribution centered at the linear function of the original forecast, with the average $a_k + b_k f_k$ and dispersion σ:

$$\Delta \mid f_k \sim N(a_k + b_k f_k, \sigma^2)$$

Values a_k and b_k are the bias-corrected terms derived from the simple linear regression of △ from f for each member of the ensemble. The approximation made is suitable for △ parameter - in our the extreme temperature.

The BMA predictive mean can be obtained using the formula:

$$M[\Delta \mid f_1, \ldots, f_K] = \sum_{k=1}^{K} w_k (a_k + b_k f_k)$$

which is a deterministic forecast, whose predictive performance can be compared with the individual forecast of the forecast ensemble or the ensemble mean.

The dispersion can be calculated using the formula given by Raftery et al. (2005):

$$D[\Delta \mid f_1, \ldots, f_K] = \sum_{k=1}^{K} w_k \left[(a_k + b_k f_k) - \sum_{l=1}^{K} w_l (a_l + b_l f_{lst}) \right]^2 + \sum_{k=1}^{K} w_k \sigma_k^2$$

where s represents space and t represents time, thus f_{lst} is forecast k in the ensemble of forecasts. The first sum represents the spread of the ensemble of forecasts, and the second sum represents the ensemble term of dispersion.

For the estimation of the Bayesian model parameters – the weights, the Maximum Likelihood Estimation (MLE) is used. In general, the method is described as follows: for $x_1, \dots, x_n$, the statistic data and θ, the unknown parameter , the likelihood function is defined as:

$$L(x_1, \dots, x_n; \theta) = \begin{cases} p(x_1, \dots, x_n; \theta) = \prod_{i=1}^{n} p(x_i; \theta), & discrete\ case \\ f(x_1, \dots, x_n; \theta) = \prod_{i=1}^{n} f(x_i; \theta), & continuos\ case \end{cases}$$

In the discrete case

$$L(x_1, \dots, x_n;\ \theta) = P_\theta (X_i = x_i, i = 1, \dots, n)$$

where $X_1, \dots, X_n$ are the observations.

In our case, the likelihood function is written as follows:

$$L(x_1, \dots, x_n;\ \theta) = \prod \mathrm{p}(\Delta | \mathrm{f}_1, \dots, \mathrm{f}_K) = \prod_{t=1}^{n} \sum_{k=1}^{K} \mathrm{w_k}\ \mathrm{g_k}(\Delta | \mathrm{f_k}, \theta_i)$$

Applying logarithm:

$$LL(\Delta | \mathrm{f}_1, \dots, \mathrm{f}_K) = \log \mathrm{L}(\Delta | \mathrm{f}_1, \dots, \mathrm{f}_K)$$

We obtain:

$$LL(\Delta | \mathrm{f}_1, \dots, \mathrm{f}_K) = \sum_{t=1}^{n} \sum_{k=1}^{51} \mathrm{w_k}\ \mathrm{g_k}(\Delta | \mathrm{f_k}, \theta_i)$$

Calculating sup $LL(\Delta | \mathrm{f}_1, \dots, \mathrm{f}_K)$ we obtain the optimization solution of Maximum Likelihood Estimation.

The algorithm **E**xpectation-**M**aximisation – **EM** is the method by which the maximum probability estimator can be obtained if the problem can be corrected. This method goes through two stages: the E stage – Expectation and the M stage – Maximization. For a parameter with a normal probability density function for which we have initialization (Paya 2005):

$$s_{kt}^{(0)} = \begin{cases} 1, if\ f_i\ is\ the\ best\ forecast \\ 0, \qquad\qquad else \end{cases}$$

Stage E: probability estimation $\mathrm{g_k}(\Delta | \mathrm{f_k}, \sigma_\mathrm{k})$,

$$s'^{(j)}_{kl} = \frac{w_k^{(j-1)} \mathrm{g_k}(\Delta_\mathrm{t} | \mathrm{f_{kt}}, \sigma_\mathrm{k}^{(\mathrm{j}-1)})}{\sum_{i=1}^{K} w_k^{(j-1)} \mathrm{g_k}(\Delta_\mathrm{t} | \mathrm{f_{kt}}, \sigma_\mathrm{k}^{(\mathrm{j}-1)})}$$

Stage M: We calculate the maximum likelihood estimator, with the estimator from stage E.

$$w_k^{(j)} = \frac{1}{n}\sum_{t=1}^{n} s'^{(j)}_{kl}$$

$$\sigma_{\mathrm{k}}^{2(\mathrm{j})} = \frac{\sum_{t=1}^{n} s'^{(j)}_{kl}\,(\Delta_{\mathrm{t}} - \mathrm{f_{kt}})^2}{\sum_{t=1}^{n} s'^{(j)}_{kl}}$$

The algorithm EM guarantees the convergence towards a local minimum. This algorithm is included in the ensembleBMA library from R software (Fraley et al. 2011). For example, for the minimum temperature, the algorithm needs approximately 27181 iterations for the 163 stations.

The calibration process has been realized with the methods developed by Fraley et al., 2011.

For this purpose, a 50 day EPS forecasts archive was created. Each day contains 51 integrations with 14 days of anticipation (highest temperatures separated from the lowest), for each of the 163 stations. This archive is modified daily by adding the observations from the previous day and the EPS forecasts of the current day, eliminating the first day from the archive. We run the programs constructed in R software daily, with the help of *"shell"* procedures for obtaining the bias coefficients and the weights of the 51 members. These apply to the current forecast for calibration.

The main stages of the operational flux are:

1. Extracting the number of test days from the archive, in order to be modeled for a specific date;
2. Fitting BMA models to ensemble forecasting data with verifying observations, with fitBMA function from R software, on the data extracted in stage 1;
3. Calibrating the current day forecast offered by MOS_EPS with anticipations of up to 15 days;
4. Creating the forecast graphic represented in *"Box-plot"* format.

The *Calibration* refers to the statistical consistency between the forecast probability distribution function and the corresponding observations (e.g. Gneiting et al. 2007).

Each box-plot is realized by emphasizing the intervals 5%-95% (quantile 5% and quantile 95%), respectively 25%-75%(quantile 25% and quantile 75%), the mean and median (quantile 50%) of the calibrated

forecast ensemble and the aberrant values within the interval 0-5% and 95-100% (Fig. 1,2).

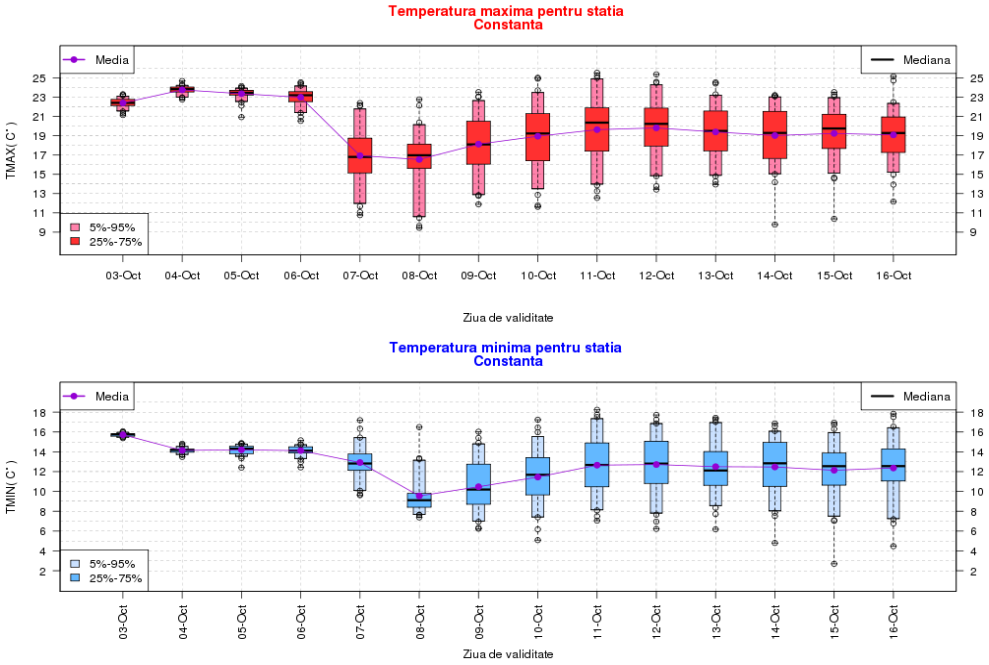


Figure 1. MOS-EPS forecast (calibrated with the BMA method), Constanţa station.

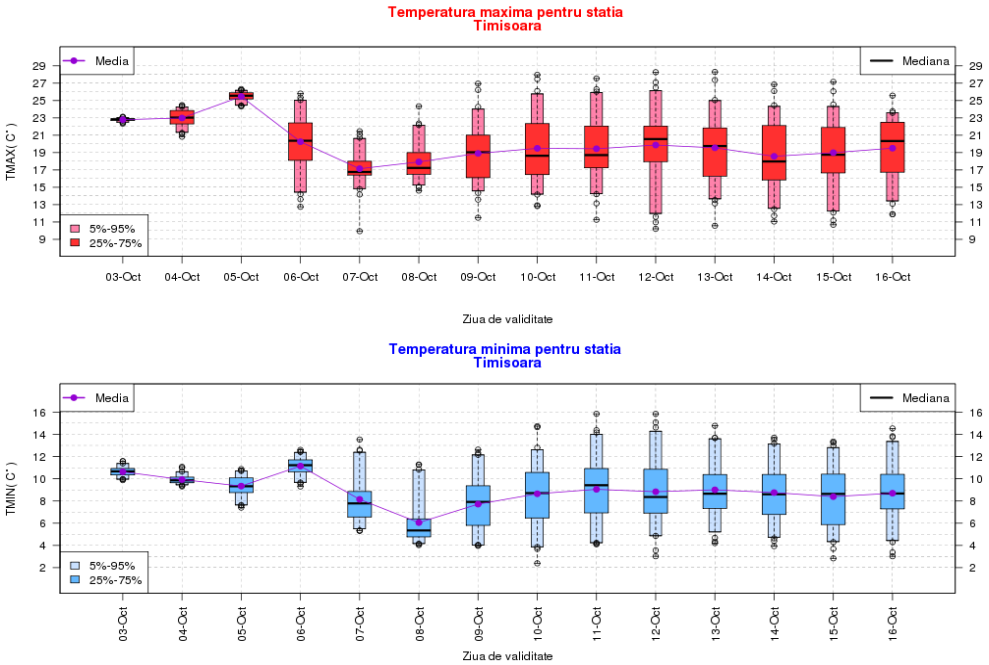


Figure 2. MOS-EPS forecast (calibrated with the BMA method), Timisoara station

This type of graphic is created daily for all 163 stations in Romania.

The verification is made by comparing the uncalibrated ensemble forecast with the calibrated one, obtained using the BMA method for maximum temperature and lowest temperature respectively in the 163 stations. In Fig. 3 we see an example of this procedure for the station in Craiova (highest temperature).

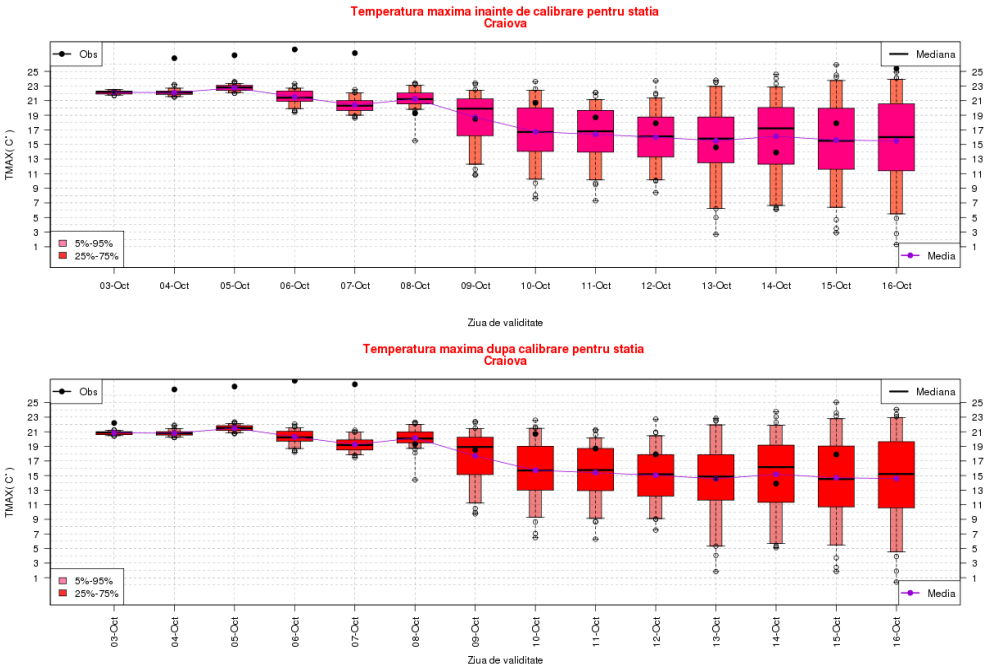
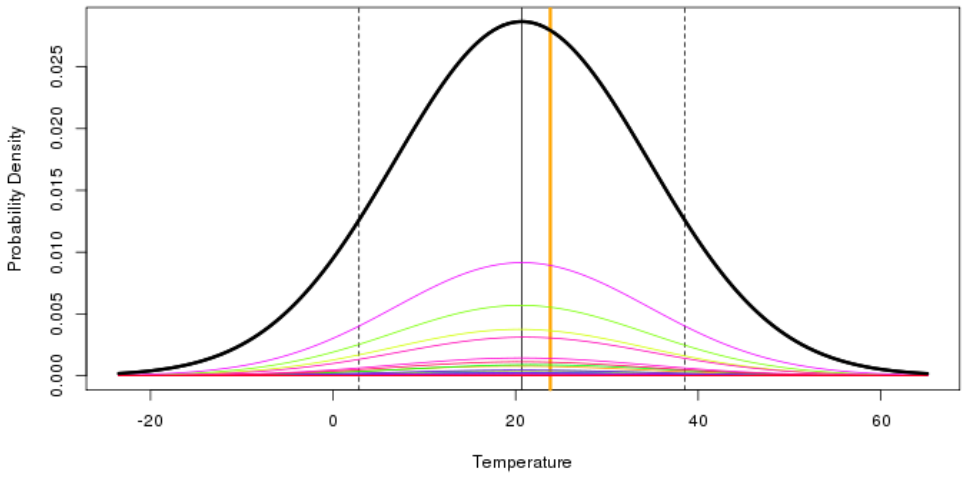


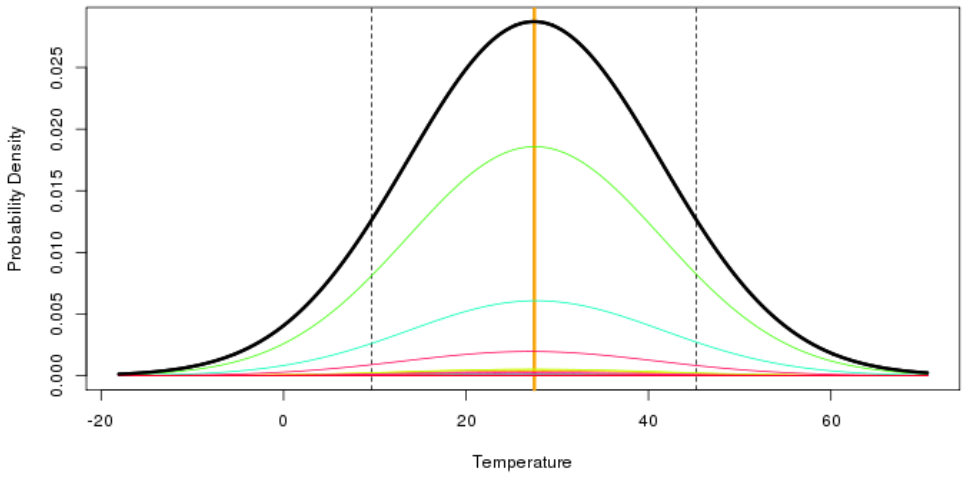
Figure 3. Maximum temperature forecast for Craiova station before and after calibration

Another testing method is based on the fact that BMA produces a predictive probability density function (PDF) for temperature, which helps offers a better view on the performance of the model. The predictive probability density function modeled by BMA is much better than the predictive probability density functions for the individual 51 members.

The graphic shows the 10^{th} and 90^{th} percentile forecast, the value of the observation in regard to each PDF- and also the PDF median offered by BMA (Fig. 4 a and b). The dashed vertical lines represent the 10^{th} and 90^{th} percentiles and the orange vertical line represents the verifying observation in relation to each predictive probability density function of the members. The thin vertical black line is the median of the BMA predictive probability density function. The black curve, shaped as a Gaussian bell curve, is the BMA PDF, and the colored curves are the weighted PDFs of the constituent ensemble members.



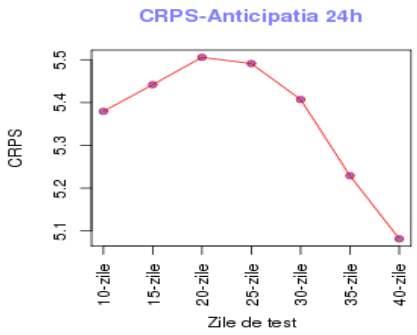
(a)



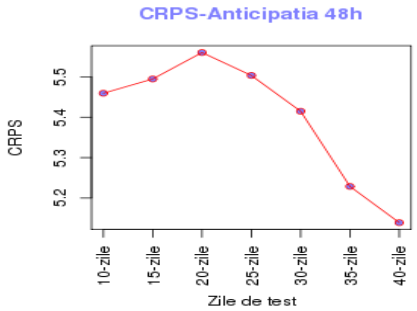
(b)

Figure 4. BMA predictive distributions for temperature valid September 22, 2012, with training period of 25 days ; a) 24 forecast hour; b) 48 forecast hour , at Craiova station.

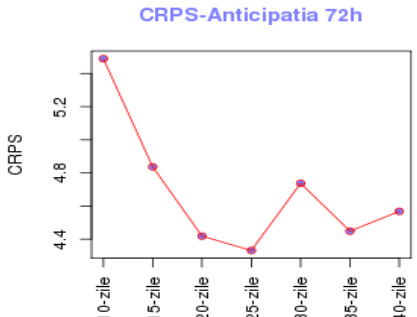
Analyzing the **CRPS** score, we see an evolution of the global skill of the ensemble forecast. CRPS measures the distance between the cumulative distribution of prediction probability and the cumulative distribution of the observation probability (CDF- cumulative density function). In figure 5 (a, b, c) the values of the CRPS score are presented, and in figure 6 (a, b, c) the MAE values.



(a)

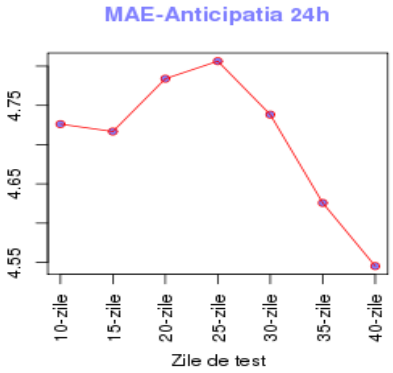


(b)

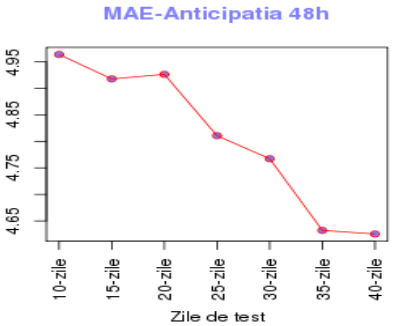


(c)

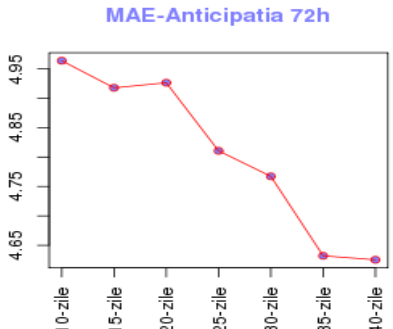
Figure 5. CRPS – anticipation: (a) 24h; (b) 48h; (c) 72h.



(a)



(b)



(c)

Figure 6. MAE – anticipation: (a) 24h; (b) 48h; (c) 72h.

3. RESULTS

The Verification rank histogram is the diagram which reflects the distribution of observations for the members of the ensemble forecast, in ascending order of the validity data, and is very useful to check the performance of the forecasts ensemble (Hamill, 2000). The Rank histogram

obtained after calibration using the BMA method for the 2nd of October 2012 (Fig. 7) shows that the BMA method makes a "cold" correction or underestimates the forecast in comparison to the verification Rank histogram of the uncalibrated forecast ensemble for the same date (Fig. 7b). We observe a slight improvement of the calibrated ensemble observation rank, meaning its slight flattening (Fig. 7 c).

After calibrating the ensemble forecast we observe that the Rank histogram is a little flatter than that of the initial ensemble forecasts, from which we can deduce that the calibrated ensemble has a better forecast for extreme temperatures than the uncalibrated ensemble.

From Table 3, we notice that in 9 of the 14 cases, the RMSE for the calibrated ensemble is lower than the RMSE for the uncalibrated ensemble; RMSE for the ensemble calibrated with BMA was lower than the RMSE for the uncalibrated MOS_EPS ensemble in 64 cases. In Table 4, we have the percentage representation of the cases in which RMSE for the calibrated ensemble is better than RMSE for each of the 51 members of the ensemble. The evaluations have been made for 14 anticipations, from 24 to 336 hours.

4. CONCLUSIONS

The paper presents the BMA method for calibrating the forecasts ensemble, as well as the stage of the operative implementation in the National Administration for Meteorology. For the study of the infusion that BMA brings to the direct ensemble proposed by MOS, we have undertaken calculus procedures of the scores in order to evaluate probabilistic forecasts. A general, practical conclusion would be that, in relation to the observations, the BMA calibration method used on MOS_EPS produces better, more reliable forecasts in comparison to those produced by each member individually. For the analyzed period, the BMA method applied to MOS_EPS tends to lower the average values of air temperature in the first 3 days, comparing it to MOS_EPS, but afterwards it estimates the extreme temperature values better than MOS_EPS. Although in the majority of the studied cases the BMA method applied to MOS_EPS has good results, there are situations when it does not estimate correctly the observation values of the respective day, as Fig. 3 shows.

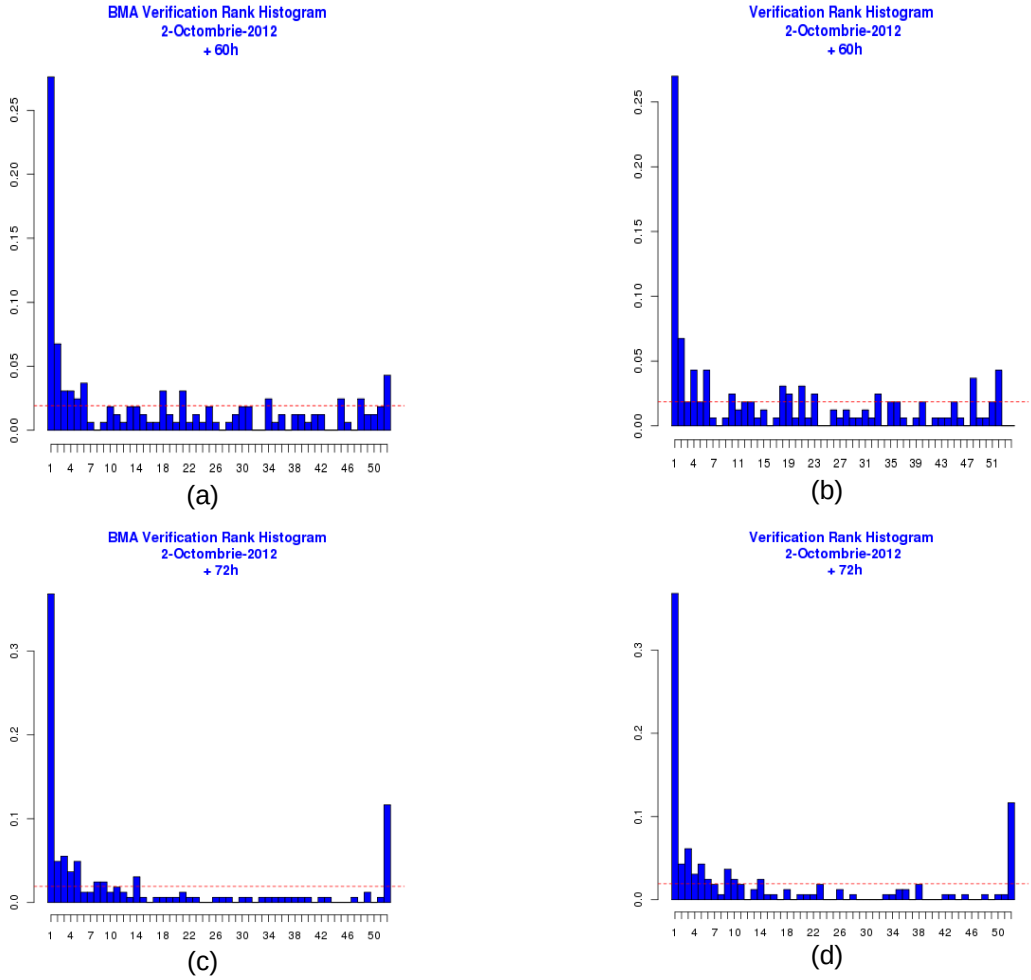


Figure 7. The Verification rank histogram for 2.10.2012. On the Ox axis we see the representation of the member's classes, and on the Oy axis we have the frequency of observations: a) calibrated BMA model – 60 h forecast; b) uncalibrated model – 60 h forecast; c) calibrated BMA model – 72 h forecast; d) uncalibrated model – 72 h forecast.

Table 4. RMSE-calibrated percentage against RMSE –uncalibrated calculated for each member for each anticipation

24h	48h	72h	96h	120h	144h	168h	192h	216h	240h	264h	288h	312h	336h
0%	70%	98%	31%	25%	49%	35%	37%	49%	35%	49%	49%	25%	31%

The probability density function for the ensemble was much better calibrated than that for each member of the ensemble.

The mean square error of the BMA calibrated ensemble was lower than the mean square error of the MOS_EPS ensemble in 64% of the studies cases, and lower than the mean square error of each member of the calibrated ensemble in 35% of the cases.

The Verification Rank Histogram is more level after calibrating the ensemble using the BMA method.

Bibliography

Azadi M, Vakili GA (2011) Calibrating surface temperature forecasts using BMA method over Iran. IACSIT Press, Singapore. 6: 23–27.

Fraley C, Raftery A, Gneiting T, McLean Sloughter J, Berrocal V (2011a): Probabilistic Weather Forecasting. The R Journal .3/1: 55–62.

Fraley C, Raftery AE, Gneiting T, McLean Sloughter J (2011b) EnsembleBMA: An R Package for Probabilistic Forecasting using Ensembles and Bayesian Model Averaging. Department of Statistic University of Washington. 1–15.

Gneiting T, Raftery AE, Westveld AH, Goldman T (2005) Calibrated probabilistic forecasting using ensemble Model Output Statistics and minimum CRPS estimation, Mon Wea Rev 133: 1098–1118.

Hamill TM (2000) Interpretation of rank histograms for verifying ensemble forecast. Mon Wea Rev 129: 550–560.

Hersbach H (2000) Decomposition of the continuous ranked probability scor for ensemble prediction systems. Weather Forecasting. 15: 559–570.

Paya M (2005) Rapport de stage: Utilisation de la BMA pour la calibration d'un systeme ensembliste de prevision adapte localement. Meteo France.

Raftery AE, Gneiting T, Balabdaoui F, Polakowski M (2005) Using Bayesian Model Averaging to Calibrate Forecast Ensembles. Mon Wea Rev 133: 1155–1173.

Roulston MS, Smith LA (2003) Combining dynamical and statis- tical ensembles, Tellus 55A, 16–30.

Vrugt JA, Clark MP, Diks CGH, Duan Q, Robinson BA (2006) Multi-objective calibration of forecast ensemble using Bayesian model averaging, Geophys Res Lett 33: 1–6.

Wang X, Bishop CH (2005): Improvement of ensemble reliability with a new dressing kernel, Q. J. Roy. Meteorol. Soc.. 131, 965– 986.

Wilks DS (2006) Statistical Methods in the Atmospheric Sciences, 2nd Edition, Academic Press. 627 pp